THE SCHOOL DISTRICT OF PHILADELPHIA

Reference Document: Assessments

Star Tests in the School District of Philadelphia: A Summary of Metrics that Describe Achievement and Growth

Ebru Erdem, Ph.D., Director, Research, Policy, and Practice; Roland Reyes, M.S., Senior Statistician

Overview

Implementing a comprehensive assessment program is a key way that school districts measure and monitor student achievement and use data to inform instructional decisions. Annual and interim assessment data from all students are critical for instruction, accountability, and research.[1] An assessment that is given to all students across all grade levels multiple times throughout the year and consistently across school years allows teachers to monitor student performance, determine any necessary additional supports for students performing below grade level, and compare student growth over time with expected growth. Such an assessment program also helps school and district leadership to monitor progress within and across years and direct resources where additional supports are needed.

In this Reference Document:

To better understand student achievement and growth across all District students, the District adopted a comprehensive assessment program that consists of universal screening and progress monitoring assessments for all grades, K-12 starting in the 2020-21 school year. In the first year of universally implemented assessments, the District used different assessment tools for K-5 and 6-12 students.[2] In the 2021-22 school year, the District began using Renaissance Star assessments as the screening and progress monitoring tool *for all grades*, assessing students during four separate time periods (called assessment windows) throughout the school year. Although all grades are required to take Star tests, the required tests for each grade vary based on the expected skills for the grade level.

There are two main types of Star tests that vary by how they are administered: Curriculum Based Measures (CBMs) and Computer Adaptive Tests (CATs) (Table 1). CBMs are 60 second assessments that measure early literacy and numeracy skills and are administered one-on-one with students, in person or virtually. CATs are administered through a computer in sessions that present the students with a fixed number of items (questions) and last 25 minutes on average. The CAT each

[1] EVERY STUDENT SUCCEEDS ACT Assessments under Title I, Part A & Title I, Part B: Summary of Final Regulations, U.S. Department of Education.

[2] In 2020-21, aimswebPlus was administered to all students in grades K-5 and Star was administered to all students in grades 6-12. Both were administered to special education students in grades 6-8.

student takes is customized; the test adapts based on the student's answers and adjusts the difficulty level of subsequent questions.

There are three kinds of Star CATs administered to SDP students: Star Early Literacy, Star Reading, and Star Math. Star Early Literacy is designed for students who do not yet read independently, and it assesses foundational reading, language and vocabulary, and numeracy skills through 27 items. Star Reading and Star Math assess, respectively, reading and math skills in multiple domains through 34 items.[3]

Star assessments provide multiple metrics that help monitor student performance at each testing window and student growth between testing windows. This reference document discusses what these metrics are, how they are calculated, and what kind of information they provide about student learning. The metrics included are: number correct (raw score), correct per minute (for CBMs), scaled score (for CATs), national percentile rank, performance tier, and student growth percentile.

Figure 1. Derivation of Star Metrics

[3] Star CATs are scored on a unified scale. That is, the score Star Early Literacy assigns to students based on the number of correct and incorrect answers and the difficulty level of the items they answered will be comparable to the score they get for the same skill level at Star Reading and Star Math. When a student scores 852 on the Star Early Literacy unified scale, they are transitioned to Star Reading and Star Math. See "Scaled Score" section below for more details.

Table 1. ELA and Math domains tested in the Star universal screening and monitoring assessments at SDP, by grade level

<table>
<tr><th rowspan="2"></th><th colspan="2">Curriculum Based Measures (CBMs)</th><th colspan="3">Computer Adaptive Tests (CATs)</th></tr>
<tr><th>ELA</th><th>Math</th><th>Star Early Literacy</th><th>Star Reading</th><th>Star Math</th></tr>
<tr><td>K</td><td>Letter Naming
Letter Sounds
Phoneme Segmentation
Receptive Nonsense Words</td><td>Numeral Recognition
Quantity Comparison</td><td rowspan="3">Word Knowledge and Skills
Comprehension Strategies and Constructing Meaning
Numbers and Operations</td><td>-</td><td>-</td></tr>
<tr><td>1</td><td>Letter Sounds
(Fall Only)
Phoneme Segmentation
Expressive Nonsense Words
Passage Oral Reading</td><td>Numeral Recognition
(Fall Only)
Quantity Comparison
Addition to 10</td><td>-</td><td>-</td></tr>
<tr><td>2</td><td>Expressive Nonsense Words
Passage Oral Reading</td><td>Addition to 10
Addition to 20
Subtraction from 10</td><td rowspan="4">Word Knowledge and Skills
Comprehension Strategies and Constructing Meaning
Analyzing Literary Text
Understanding Author's Craft
Analyzing Argument and Evaluating Text</td><td rowspan="4">Numbers and Operations
Algebra
Geometry and Measurements
Data Analysis, Statistics, and Probability</td></tr>
<tr><td>3</td><td>Passage Oral Reading</td><td>Subtraction from 10
Mixed Addition and Subtraction
Multiplication to 100</td><td rowspan="3">Teachers <u>may</u> decide to administer Star Early Literacy for students with scaled scores below 852 <u>in addition</u> to Star Reading and Math</td></tr>
<tr><td>4</td><td>Passage Oral Reading</td><td rowspan="2">-</td></tr>
<tr><td>5</td><td>Passage Oral Reading</td></tr>
<tr><td>6-12</td><td>-</td><td>-</td><td></td><td></td><td></td></tr>
</table>

Source: Technical manuals for Star CBM Reading (pp.14-19) and Math (p.4), Star Early Literacy (pp.10-18), Star Reading (pp.15-16) and Math (p. 15).[4]

[4] Star CBM Reading Technical Manual 2021:
https://help2.renaissance.com/US/PDF/starcbm/StarCBMReadingTechnicalManual.pdf
Star CBM Math Technical Manual 2021:
https://help2.renaissance.com/US/PDF/starcbm/StarCBMMathTechnicalManual.pdf
Star Assessments for Early Literacy Technical Manual 2022:
https://help2.renaissance.com/US/PDF/SEL/SELRPTechnicalManual.pdf
Star Assessments for Reading Technical Manual 2022:
https://help2.renaissance.com/US/PDF/SR/SRRPTechnicalManual.pdf

Star Raw Scores (Number Correct)

Star Curriculum Based Measures (CBMs), administered one-on-one, measure how many items a student answers correctly in one minute, and this is reported in the number correct raw score. For example, in the Letter Naming CBM, a kindergartener will be shown letters during the span of a minute, and the raw score will be the number of letters they can identify correctly.

Each of the Curriculum Based Measures are assessed through multiple forms; for example, various texts might be used for Passage Oral Reading. However, the difficulty of these forms varies; for example, one Passage Oral Reading text might have more difficult vocabulary than another one. Star performs an equating process on the raw scores to adjust for the difficulty of different forms and the resulting equated score is called *Correct Per Minute* (CPM).[5] CPM scores can be interpreted on the same scale, are comparable across students, and form the basis for norm-referenced scores such as National Percentile Rank (see below).[6]

In Star Computer Adaptive Tests (CATs), the raw score is again based on the number of correctly answered items (out of 27 in Star Early Literacy and 34 in Star Reading and Math); however, the CAT items in each test have different levels of difficulty for each student because of the adaptive nature of the test. For example, two students who answered 20 correct out of the 34 Star Reading items might have different performance levels because one of the students received more difficult items due to the adapting. Because teachers do not know the difficulty level of the tests that different students take, the raw score in the CAT (number correct) is not very informative for teachers. Instead, Star uses scaled scores that take into account the difficulty of the items answered.[7,8] Scaled scores are on a unified scale, which means that they can be compared across students and grade levels, and over time.

Star Assessments for Math Technical Manual 2021:
https://help2.renaissance.com/US/PDF/SM/SMRPTechnicalManual.pdf

[5] Star CBM uses the circle-arc equating method to link all forms to the base form, the form identified as the easiest. An equated correct count is calculated for each form. The equated correct count multiplied by 60 and divided by the total time the student took to complete the assessment is recorded as the Correct per Minute (CPM). Passage Oral Reading equating has not been completed yet for grades 4 to 6, benchmarks and percentile ranks for these grades are based on mean performance on three passages in each window. For details, see Star CBM Reading Technical Manual pp. 22-26 and https://help2.renaissance.com/starcbm/CPMScores

[6] CPM is not a scaled score as the scaled scores are calculated from the CAT raw scores (see below). This is because scales for each Curriculum Based Measure are different. Renaissance refers to CPM as an "adjusted" score.

[7] Original Star Reading item calibration was done in Spring 1998, and Star Math item calibration was done in Spring 2001. For both tests, new items are added to the item banks after going through "dynamic calibration," after online data collection.

[8] The unified scales for Star CAT were developed based on Item Response Theory (IRT) and use a Rasch IRT model.

Star Scaled Scores

The Star scaled score is a student performance metric on the Star Computer Adaptive Tests (CATs). Similar to a raw score (number correct), the scaled score provides information about how a student performed on the test. The scaled score, however, differs from the raw score because it is calculated based on both the number of correct answers and the difficulty of the items.[9] This improves interpretability of student performance over raw scores when different students receive different sets of items that vary in difficulty.

Star determines the difficulty level of its CAT questions (items) through item calibration. Star analyzes how students with known performance levels answer each item. If an item is answered correctly by a low percentage of students, it is deemed a difficult item that requires a higher level of skill. In this sense, the scaled score, through item-difficulty determination, implicitly includes a comparison of test-takers to their peers. Difficulty of each item, along with the pattern of correct and incorrect answers, determines the scaled score.

Star uses a unified scale,[10] which ranges from 200 to 1400, to report scaled scores on all CATs across all grade levels, K-12 (Table 2). This means that Star scaled scores can be used for comparisons across students, across grade levels, and notably, across Star Early Literacy and Star Reading and Math. The unified scale allows us to do multiple analyses:

- Comparing performance of different students within the same testing window
- Tracking student performance over time as they progress into higher grades
- Transitioning students from assessments designed for emergent readers to assessments designed for independent readers while continuing to track their abilities continuously as the assessment tool for their reading skills changes

Table 2. Scaled score ranges for Star Computer Adaptive Tests

Star Computer Adaptive Test	SDP Recommended Grade Span	Scaled Score Range
Early Literacy	K-2	200-1100
Reading	3-12	600-1400
Math	3-12	600-1400
Early Literacy Spanish	K-2	0-1100
Reading Spanish	3-12	600-1400
Math Spanish	3-12	600-1400

[9] Star scaled scores are different from Pennsylvania System of School Assessment (PSSA) scale scores because the PSSA is a standards-based, criterion-referenced test, not norm-referenced. The PSSA scale scores range from 600 to 1600 for each grade, with 1000 as the proficient cut point in every grade, and they cannot be compared between years for a particular student. PSSA scale score ranges: https://www.stateboard.education.pa.gov/Documents/About%20the%20Board/Board%20Actions/2015/PSSA%20Cut%20Scores.pdf

[10] Primer on Star scores from Renaissance Learning: https://help2.renaissance.com/goals
Renaissance Learning. (2021). Star Unified Scale: https://renaissance.widen.net/s/w6p9f5pcpm/r63395

As Table 2 shows, the ranges of Star scaled scores span across grade levels from grades K-12. To understand where a student's performance stands compared to their peers at the same grade level, norm-referenced metrics are necessary. Scaled scores are the basis of norm-referenced metrics such as National Percentile Rank and Performance Tiers for Star CATs.

National Percentile Rank

National Percentile Rank (NPR) compares each student's performance to a national sample and shows what percent of the students in the national norm sample received a scaled score lower than or equal to the student's score.[11] A *norm sample* is a nationally representative sample of U.S. students across all grade levels.[12] Scaled scores and the distribution of the scaled scores of the students in this sample were used to determine the expected performance of U.S. students at each grade level. After developing the unified scale, Star developed new norms in 2017 by mapping the scaled scores on the unified scale to the performance and distribution of scores observed for the norm sample for fall and spring of that year.

The norms show the expected scaled scores at a given grade level in the beginning and at the end of the year, and the National Percentile Rank shows how a student performs compared to the national norm (See Tables 4 and 5 below). For example, a sixth-grade student who receives a scaled score of 1052 in the fall Star Reading assessment would have scored better than 40% of all sixth-graders (and is in the 40th NPR). Compare this to a sixth-grade student who receives the same scaled score, 1052, in the spring Star Reading assessment; this student would have scored better than only 30% of all sixth-graders.

National Percentile Rank is useful in understanding the performance level of a student compared to the expected performance appropriate for the student's grade level. However, it is not a very actionable metric without additional information, such as benchmarks of expected performance. Benchmarking is the process of mapping performance on Star tests to expected performance on state standardized tests. Districts use benchmarking to identify students who need additional supports to move toward a particular performance level, most often proficient or advanced at state standardized tests. At SDP, we use Performance Tiers for this purpose.

[11] A similar metric that Star also reports is the Normal Curve Equivalent (NCE). NPR is an ordinal scale; that is a 1 percentile increase spans across a wider range of scaled scores in the upper and lower ends of the 1-99 NPR range, making operations like averaging incorrect. In comparison, NCE is on an interval scale; that is to say, a 1 point increase in NCE spans across the same range of scaled scores at any point in the 1-99 range. This metric is primarily used for aggregations and statistical operations.

[12] Star conducted a study to gather fall and spring norming samples of test-takers between August 15, 2014, and June 30, 2015. The demographic and geographic characteristics of the sample are provided in the Norming/Sample Characteristics of the Star Reading and Math Technical manuals, linked above in fn. 4.

Performance Tiers

Star has determined that students "whose test scores place them in a National Percentile Rank of 40 or higher (ELA) or 70 or higher (Math) will likely meet end-of-year performance goals as defined by the state and local standards."[13] SDP uses the following Star Reading, Star Math, and Star Early Literacy cutoffs for determining the performance tiers and implements interventions to provide the necessary supports to students who are below benchmark (Table 3).

Table 3. Performance tiers at the SDP

Performance Level	Description	NPR Range (Star Reading and Star Early Literacy)	NPR Range (Star Math)
At/Above Benchmark	Students meeting or exceeding the benchmark score	≥ 40	≥ 70
On Watch	Students slightly below the benchmark score	25-39	25-69
Strategic Intervention	Students below the benchmark score	10-24	10-24
Intensive Intervention	Students far below the benchmark score	<10	<10

Source: SDP Office of Assessments and Renaissance-Defining Benchmarks in Star Assessments, https://doc.renlearn.com/KMNet/R62855.pdf

Once a district sets the NPR cutoffs for performance levels, they do not change at different testing windows or across grade levels. For example, for Star Reading and Star Early literacy, at or above 40th percentile is considered to be At/Above Benchmark (Tier 1) for all grade levels and for all testing windows. However, the scaled score ranges that are associated with these static NPR cutoffs change from grade to grade and between SDP's fall, winter 1, winter 2, and spring testing windows within the same grade. For example, the third-grade Star Reading scaled score cutoffs associated with the NPR benchmarks for the four performance tiers (10th, 25th and 40th percentiles) are respectively 864, 908, and 938 in September and increase as the year progresses, ending with, respectively, 901, 942, and 969 (Figure 2). Scaled score benchmark cutoffs for each performance tier are used in all grades for Star Reading and Star Early Literacy (Table 3) and for Star Math (Table 4).

[13] Renaissance Learning. (2021). Star Reading Unified Scale Benchmark Cut Scores: https://help2.renaissance.com/US/PDF/SR/SRUnifiedBenchmarksCutScores.pdf
Renaissance Learning. (2021). Star Math Unified Scale Benchmark Cut Scores: https://help2.renaissance.com/US/PDF/SM/SMUnifiedBenchmarksCutScores.pdf (Note that, since the 2021-22 school year, SDP benchmark for At/Above Benchmark in Math has been 70th percentile, not 40th.)

Figure 2. Third grade Star Reading Unified Scale benchmark cutoffs for each performance level from September to May

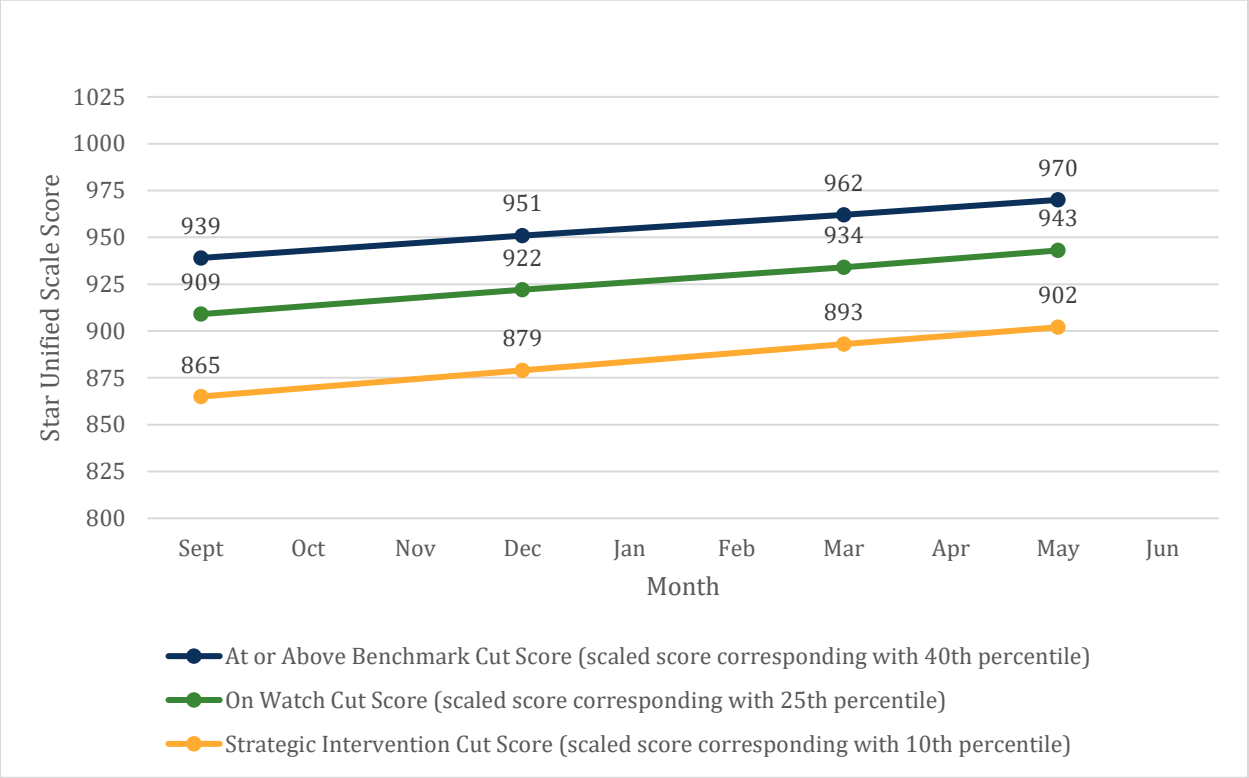

Table 3. Reading benchmark cutoffs used by SDP

		Fall (September)	Winter 1 (December)	Winter 2 (March)	Spring (May)
Grade	**NPR**	**Scaled Score**	**Scaled Score**	**Scaled Score**	**Scaled Score**
K	10	622	655	687	709
	25	662	692	722	742
	40	690	720	749	769
1	10	691	717	743	760
	25	730	756	783	800
	40	752	781	809	828
2	10	794	814	833	846
	25	835	855	876	889
	40	868	887	905	918
3	10	865	879	893	902
	25	909	922	934	943
	40	939	951	962	970
4	10	913	922	931	937
	25	954	963	971	977
	40	982	991	999	1005
5	10	949	957	965	970
	25	993	1000	1007	1012
	40	1021	1028	1035	1040
6	10	982	988	995	999
	25	1025	1031	1038	1042
	40	1052	1059	1065	1070
7	10	1002	1007	1012	1015
	25	1045	1050	1056	1059
	40	1073	1079	1085	1089
8	10	1022	1027	1032	1035
	25	1066	1071	1077	1080
	40	1096	1101	1107	1110
9	10	1042	1047	1053	1056
	25	1086	1090	1094	1097
	40	1115	1119	1122	1125
10	10	1058	1058	1059	1059
	25	1098	1100	1101	1102
	40	1126	1128	1130	1131
11	10	1058	1061	1065	1067
	25	1102	1105	1108	1110
	40	1131	1134	1137	1139
12	10	1066	1069	1071	1073
	25	1114	1115	1117	1118
	40	1145	1146	1147	1148

Source: https://help2.renaissance.com/US/PDF/SR/SRUnifiedBenchmarksCutScores.pdf and documentation provided by Renaissance to SDP. SDP tests in four windows rather than three; Winter 1 and Winter 2 scaled score cutoffs listed here correspond to the mid-point of the respective window.

Table 4. Math benchmark cutoffs used by SDP

		Fall (September)	Winter 1 (December)	Winter 2 (March)	Spring (May)
Grade	**NPR**	**Scaled Score**	**Scaled Score**	**Scaled Score**	**Scaled Score**
1	10	700	727	753	771
	25	733	760	787	805
	70	802	828	854	871
2	10	792	814	835	849
	25	829	848	867	879
	70	888	907	926	938
3	10	849	865	882	893
	25	882	901	919	932
	70	941	961	981	994
4	10	894	908	923	932
	25	934	947	961	970
	70	996	1011	1025	1035
5	10	932	944	956	964
	25	972	984	996	1005
	70	1040	1051	1063	1070
6	10	966	978	990	997
	25	1010	1019	1029	1035
	70	1085	1093	1101	1107
7	10	984	992	1000	1005
	25	1031	1037	1043	1047
	70	1109	1115	1121	1125
8	10	995	1002	1010	1015
	25	1046	1052	1059	1063
	70	1131	1137	1144	1148
9	10	1003	1011	1019	1025
	25	1053	1059	1064	1068
	70	1134	1139	1144	1147
10	10	1000	1008	1016	1022
	25	1053	1059	1064	1068
	70	1136	1141	1147	1150
11	10	1028	1031	1033	1035
	25	1071	1074	1077	1079
	70	1150	1153	1156	1159
12	10	1035	1037	1039	1040
	25	1081	1083	1086	1087
	70	1160	1162	1163	1165

Source: https://help2.renaissance.com/US/PDF/SM/SMUnifiedBenchmarksCutScores.pdf and documentation provided by Renaissance to SDP. SDP tests in four windows rather than three; Winter 1 and Winter 2 scaled score cutoffs listed here correspond to the mid-point of the respective window.

Student Growth Percentile

Student Growth Percentile (SGP) is a metric that describes a student's growth on the Star CAT relative to their academic peers nationwide. In this context, "academic peers" is defined as the group of grade-level students who received similar scaled scores on prior administrations of Star.[14] SGP is a norm-referenced measure that ranges from 1-99 and is interpreted like a percentile rank. For example, an SGP of 70 means a student grew equal to or more than 70% of their grade-level peers with similar past scaled scores nationwide. For context, an SGP of 50 indicates typical growth with respect to the student's grade-level peers with similar past scaled scores.

SGPs are estimated using a statistical technique called quantile regression. In Star, at least two scores from different testing windows are required to calculate an SGP: the student's most recent score within the past 18 months and a score from a previous testing window.[15] It is important to note that the Star testing windows used for SGP calculations differ somewhat from the District's testing windows. Specifically, Star's SGP windows are:

- Fall (August 1 – November 30)
- Winter (December 1 – March 31)
- Spring (April 1 – July 31)

When multiple tests have been completed within a Star SGP window, certain tests will be prioritized when calculating SGP:

- Fall: First test taken
- Winter: Test closest to January 15
- Spring: Last test taken

One property of SGP is that it is possible for students to score an SGP of 1 to 99 regardless of prior performance. While SGP helps contextualize student growth, it is important to keep in mind that they are interpreted relative to the student's academic peer group (i.e. grade-level students with similar prior scaled scores) and do not describe growth on an absolute scale.[16,17]

Renaissance maintains growth norms used for calculating SGP. SGP is currently available for Star Early Literacy (K-3), Star Reading (1-12), Star Math (1-12), Star Early Literacy Spanish (K-3), Star Reading Spanish and Star Math Spanish (1-8).

[14] Renaissance Learning. (2021). Student growth percentile in Star Assessments: https://doc.renlearn.com/KMNet/R00571375CF86BBF.pdf

[15] If available, the Star statistical model will incorporate an additional pretest score from the previous year when calculating SGPs. This applies to English-language versions of the tests only at the time of this writing.

[16] Castellano, K. E., & Ho, A. D. (2013). A practitioner's guide to growth models: https://files.eric.ed.gov/fulltext/ED551292.pdf

[17] SGP is distributed uniformly across the 1-99 range; that is, each student has the same possibility of being at any growth level regardless of their performance level. At the individual level, practitioners should use performance (e.g. NPR) and growth (SGP) metrics together to have a full picture of how individual students are doing compared to their peers. Aggregating SGP (e.g. mean, median) should be done with caution because as larger numbers of student SGPs get aggregated the uniform distribution reverts into a bell curve (normal distribution).